(wileyonlinelibrary.com) DOI: 10.1002/pst.1592

Published online 30 August 2013 in Wiley Online Library

Guidance on the implementation and reporting of a drug safety Bayesian network meta-analysis

David Ohlssen,[a]* Karen L. Price,[b] H. Amy Xia,[c] Hwanhee Hong,[d] Jouni Kerman,[e] Haoda Fu,[b] George Quartey,[f] Cory R. Heilmann,[b] Haijun Ma,[c] and Bradley P. Carlin[d]

The Drug Information Association Bayesian Scientific Working Group (BSWG) was formed in 2011 with a vision to ensure that Bayesian methods are well understood and broadly utilized for design and analysis and throughout the medical product development process, and to improve industrial, regulatory, and economic decision making. The group, composed of individuals from academia, industry, and regulatory, has as its mission to facilitate the appropriate use and contribute to the progress of Bayesian methodology. In this paper, the safety sub-team of the BSWG explores the use of Bayesian methods when applied to drug safety meta-analysis and network meta-analysis. Guidance is presented on the conduct and reporting of such analyses. We also discuss different structural model assumptions and provide discussion on prior specification. The work is illustrated through a case study involving a network meta-analysis related to the cardiovascular safety of non-steroidal anti-inflammatory drugs. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: drug safety meta-analysis; rare events; mixed treatment comparisons; network meta-analysis; multiple outcomes; reporting Bayesian analysis; prior sensitivity

1. INTRODUCTION

Meta-analysis has been defined as ‘a statistical analysis which combines the results of several independent studies considered by the analyst to be *combinable*’ [1]. When combining information from different trials, the following factors must be carefully considered: (1) what data to combine; (2) the assumptions or models used to describe the relationship between the trial-specific parameters and the potential covariates (trial or subject-specific); (3) how to carry out statistical inference, that is, to quantify the information about the parameter(s) of interest; and (4) drawing conclusions from the results. Steps (2) and (3) typically require specialist statistical skills in addition to some level of subject-matter expertise. Within the context of drug safety evaluation, based on combining information from multiple randomized control trials, in this paper, we shall focus on steps (2) and (3) in a Bayesian framework.

In terms of drug safety, there have been a number of high profile examples in which a published meta-analysis has raised concerns, leading to product withdrawals or large reductions in usage of the therapy under investigation (e.g., [2, 3]). A running example, considered throughout our presentation, is an investigation involving cardiovascular outcomes related to non-steroidal anti-inflammatory drugs (NSAIDs) [4]. By combining direct and indirect evidence from multiple treatments, the authors go beyond classical meta-analysis, which only considers direct evidence. This requires methods for *network meta-analysis* (NMA) [5] or *mixed treatment comparison* (MTC) models [6], which are typically implemented in practice using a Bayesian statistical framework [7, 8] (although classical modeling approaches, such as those presented in [5], are possible).

Bayesian methods are based upon the interpretation of probability as a (possibly subjective) measure about one’s current state of knowledge. This allows the expression of uncertainty about any unknown quantity and differs markedly from the classical, long-run ‘frequentist’ view, which restricts probability statements to events that can be replicated. When applying Bayesian ideas to statistical modeling, we must begin with a *prior distribution* that embodies any existing prior knowledge or beliefs concerning the set of model parameters. Once further information (say, study data) is obtained, the prior is updated to the *posterior distribution* using the basic principles of conditional probability. Statistical inferences (point estimates, confidence intervals,

[a] Novartis Pharmaceuticals Corporation, East Hanover, NJ 07936, USA

[b] Eli Lilly and Company, Indianapolis, IN 46285, USA

[c] Amgen, Inc, Thousand Oaks, CA, USA

[d] Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA

[e] Novartis Pharma AG, Basel, Switzerland

[f] Genentech, Inc, South San Francisco, CA, USA

*Correspondence to: David Ohlssen, Integrated information Science, Novartis Pharmaceuticals, One Health Plaza, East Hanover, NJ 07936, USA.
E-mail: david.ohlssen@novartis.com

and hypothesis tests) then emerge straightforwardly from appropriate summaries of the posterior; see, for example, [9] for a full description of Bayesian modeling and computational tools.

To practitioners, the fundamental philosophical differences between the Bayesian and the frequentist framework are often of little concern. Rather, the key point is how Bayesian methods can be used to add value when tackling a particular problem. When conducting a meta-analysis, a variety of potential advantages of the Bayesian approach are given in an insightful recent review [10]. However, these authors acknowledge that many of these points can also be tackled from frequentist perspective. In the literature, there is often a tendency to compare a relatively naive classical approach with a more tailored Bayesian approach or vice versa. Therefore, in the context of rare adverse event meta-analysis and evidence synthesis, in the subsequent text, we emphasize later three key advantages of applying a Bayesian analysis that might be used in conjunction with classical meta-analysis techniques, along with some commentary on how the Bayesian approach adds value.

First, Bayesian methods offer unified modeling, as well as the ability to explore a wide range of modeling structures that permit the synthesis of evidence from multiple sources and may also incorporate multiple treatments. For instance, with appropriate prior specification, models with range assumptions, such as those described in [11], can be easily fitted in the Bayesian framework using MCMC computational methods, facilitating sensitivity analysis across a range of potential structures. Model comparison can also be accomplished using the deviance information criterion (DIC) [12] or other straightforwardly computed and interpreted measures [13]; model averaging is also possible [14]. Additionally, Bayesian simulation-based methods do not need to rely on asymptotic approximations and can straightforwardly accommodate uncertainty about nuisance parameters and missing data, often leading to more accurate statements about uncertainty in the overall conclusions.

Second, Bayesian methods permit formal incorporation of other sources of evidence by utilizing prior distributions for model unknowns. When considering meta-analysis of rare events, the use of prior distributions not only provides potential benefits but may also lead to concerns regarding the impact of inappropriately informative priors. For example, while one potential advantage of the Bayesian framework is the ability to incorporate prior information regarding background rates of events, there is limited research into the appropriate way to accomplish this in practice. Sensitivity analysis is typically required, along with careful explanations of how all prior distributions were formed. A second example of the potential benefit here is the ability to adopt prior distributions on parameters expressing between-study variability in models that incorporate random effects. The use of weakly informative priors can make such models much easier to identify, and avoid some of the optimization problems associated with classical estimation techniques.

Finally, Bayesian methods enable the user to express analysis results in terms of direct probability statements about all model unknowns, given all available evidence. Compared to classical *p*-values (which have a somewhat counterintuitive interpretation as the long-run probability of obtaining findings as surprising as those observed or more so had the null hypothesis been true), such Bayesian probability statements (say, the posterior probability that a certain drug is safest) are often much more easily understood by practitioners, ultimately improving the decisions that they make.

In the remainder of this paper, we will provide guidance on the methods, implementation, and reporting of a Bayesian NMA in the context of drug safety assessment as follows. First, in Section 2, we provide a review of Bayesian meta-analysis and NMA; Section 3 then provides full detail on the models and framework that will be used throughout. Section 4 discusses a strategy and checklist for the development and reporting of a Bayesian meta-analysis in the context of drug safety. Section 5 illustrates in the context of our NSAIDs example, utilizing the models and methods described in Section 3. Finally, Section 6 discusses our findings and offers areas for further exploration.

2. A REVIEW OF BAYESIAN META-ANALYSIS

2.1. A brief review of methods for Bayesian meta-analysis

In the case of a comparison involving two treatments with a single primary outcome, the classical texts on Bayesian meta-analysis utilized Monte Carlo methods [15] and MCMC methods [16], in a model with random effects placed on the treatment contrasts. Since these early developments, a vast literature has evolved with particular emphasis on random effects models (for a recent extensive review, see [17]), including meta regression [18], alternative scales and link function [19], selection of prior distributions for variance components [20, 21], combining individual patient data with aggregate data [22], combining data with different levels of rigor and relevance via prior elicitation [23], flexible random effects assumptions [24, 25], and subgroup analysis [26]. Note that all of this work focuses on approaches to deal with heterogeneity, either indirectly through modeling assumptions or more directly by introducing additional information together with an appropriate regression structure. The emphasis on Bayesian random effects models is likely due to the potential advantages of MCMC in a generalized linear mixed model (GLMM) setting over alternative classical estimation procedures (e.g., see [27]). Note that these developments are applicable to all forms of likelihood (e.g., binomial, normal, and Poisson) and link functions within a GLMM framework. In the following case study, we focus on a Poisson sampling model with a log link. In terms of guidance on implementation of Bayesian meta-analysis for the practicing statistician, Dias *et al.* [7] and Spiegelhalter *et al.* [10] both provide excellent tutorial-style material, along with numerous examples and guidance on software implementation.

As previously mentioned, Higgins *et al.* [17] provides a recent review of random effects meta-analysis from both classical and Bayesian perspectives. When comparing the two approaches, they emphasize that a Bayesian approach has the additional advantages of flexibility, allowing incorporation of full uncertainty in all parameters (but not uncertainty in the model) and of yielding more readily interpretable inferences. Further, they note that in cases with a large number of large studies, Bayesian nonparametric random effects approaches, such as those discussed in [25], can be considered as an alternative to the typical normal random effects model. In more typical situations, the nonparametric model might not be well identified. Therefore, sensitivity analysis with a range of parametric assumptions, such as *t*-distributions and skewed distributions, could be examined using the DIC. In cases with a few studies, even a normal random effects model can be difficult to fit using classical estimation techniques. However, in the Bayesian framework, because informative prior distributions for the extent of heterogeneity can be utilized, such models can be fitted with careful prior sensitivity analysis.

Looking beyond meta-analysis involving two treatments with a single primary outcome, further work has developed involving multiple treatments [5,6] and multiple outcomes [28–30]. As mentioned in Section 1, the former has been referred to as NMA or MTCs. The latter is often described under the umbrella of multivariate meta-analysis; see [31] for a recent review focusing on the Bayesian perspective. Because of the ability to estimate parameters with greater precision due to in-built borrowing of strength, both approaches could be particularly valuable when dealing with rare safety events. In this paper, we shall focus on a framework that accommodates multiple treatments within an NMA setting. However, it should be emphasized that the work presented in Section 3 is applicable to the case involving two treatments and a single outcome, as discussed earlier in this section. In addition, some possible extensions to multiple outcomes will also be examined in Subsection 3.3. This work too is directly applicable to a situation involving just two treatments. NMA is able to estimate more parameters and with greater precision because of its borrowing of strength, particularly in the case of sparse data.

2.2. An overview of Bayesian network meta-analysis for safety events

In this section, we focus on Bayesian NMA models that will be applied to the case study involving an NMA of NSAIDs. The most commonly used models will be emphasized, together with extensions and alternatives that are particularly relevant when dealing with rare safety events.

The analysis of the NSAID data presented in [4] that utilized a Bayesian random effects NMA, with a Poisson sampling model, following a structure that has been proposed in number of recent guidance documents on best practices [7, 8]. This approach extends earlier work on Bayesian meta-analysis involving just two treatments [16]. In the subsequent text, this model shall be referred to as the Lu and Ades (LA) random effects model, after the classical texts [6, 32, 33] from which the models described in [7, 8] are drawn. It should be noted that Lu and Ades [6, 32, 33] introduce a broader class of models and in each paper utilize slightly different parameterizations. The parameterization we shall subsequently utilize for the LA model will follow Section 5 of [7, 8].

A key feature of this approach is that a separate fixed effect (or *baseline parameter*) is associated with each trial, leading to a structure where there is no borrowing of strength among the set of trial baseline parameters. This requires that a concurrent control treatment be assigned in each trial, thus avoiding cross-trial bias (the potential bias caused by borrowing strength across trial baseline parameters) [11]. In the case of the treatment effects or treatment contrasts, within the linear predictor, exchangeability is assumed through normally distributed random effects. When assuming no variation among the set of treatment contrasts across trials, the LA fixed effects NMA model, emphasized by Dias *et al.* [7] and Hoaglin *et al.* [8], is formed. These two models will be the focus of Section 3.1.

One of the potential challenges to dealing with meta-analysis of rare events is the limited information provided by each trial. It has been noted that this can lead to challenges when estimating the LA random effects model (see [7], p. 40). In classical meta-analysis, conducted on a relative scale, it is a common practice to remove studies in which there are no events in each of the treatment arms from a meta-analysis (note that this is not the case for the risk and rate difference scales). A rationale for dropping these studies is that from a likelihood perspective, they provide no information on the magnitude of a relative treatment effect [34]. Similarly, in the Bayesian setting, when using the modeling structure adopted in both LA random and fixed effects model, with priors that aim to be non-informative, these trials will contribute very little and can also lead to numerical instability [35]. If a continuity correction is applied, then these studies will once again contribute. However, in a meta-analysis involving rare binary events, it is acknowledged that continuity corrections can be very influential and their use can produce highly biased results and misleading conclusions [35, 36].

In the Bayesian setting, a more principled alternative to continuity corrections is the use of informative priors on background rates or control rates. However, as noted in [37], the use of informative priors in NMA is dependent on the choice of parametrization. Therefore, in Section 3.2, a way to parametrize the LA random effects model is explored where the same reference treatment is used in every trial, making the formulation of informative priors on control rates more feasible.

Along with informative priors on background rates, a further possibility is to form a random effects model that borrows strength across baselines. This can be achieved by either including the baseline parameters as a set of random effects (see models SST2-HOM, SST2-HET, and SST-3 in [6]) or including all treatment arms as random effects to form a multivariate meta-analysis [38, 39]. Because of the potential for cross-level bias, caused when the estimation procedure allows borrowing of strength across control groups and potentially introduces bias by not associating a separate concurrent control with each study, such model assumptions must be carefully examined. However, as acknowledged in [7], such an approach might be worthy of consideration in a case with many zero cells, such as rare event meta-analysis ([7], pp. 40–41). We return to these ideas in greater detail in Subsection 3.3.

3. METHODS

3.1. Bayesian fixed and random effects models for network meta-analysis

To provide a framework, NSAIDs context, and corresponding notation that will be used throughout, let Y_{ik} represent the number of cardiovascular events in arm k of NSAIDs trial i during the planned trial follow-up period, E_{ik} the exposure time in person-years, and λ_{ik} the event rate. We then assume that the Y_{ik} are conditionally independent with distribution

$$Y_{ik} \sim \text{Poisson}(E_{ik}\lambda_{ik}),\ i = 1,\ldots,N,\ k = 1,\ldots,K. \tag{1}$$

To form the random effects structure, a baseline treatment b must be the specified for each of the trials. By taking the log link function $\eta_{ik} = \log(\lambda_{ik})$, the LA model assumes

$$\eta_{ik} = \begin{cases} \mu_i & k = b \\ \mu_i + \delta_{ibk} & k > b \end{cases} \tag{2}$$

$$\text{and}\quad \delta_{ibk} \overset{ind}{\sim} N\left(d_{1k} - d_{1b}, \sigma^2\right), \tag{3}$$

where η_{ik} is a continuous measure of the treatment effect in arm k of trial i (in this case, a log rate), μ_i is the effect of baseline treatment b in trial i, δ_{ibk} is the trial-specific treatment effect of treatment k relative to treatment b, and σ^2 represents the variance of the random effects, assumed to be homogenous across all treatment comparisons. An alternative is the corresponding fixed effects model, which assumes no variation among the treatment effects across trials, so $\delta_{ibk} \equiv d_{1k} - d_{1b}$. It is also possible to relax the assumption of a homogenous variance components [6, 33]. In all of these cases, models (2)–(3) are called *contrast-based* (CB) models because the treatment contrasts (δ_{ibk} in the random effects case) are the basic units being modeled and across which exchangeability is being assumed.

In (2), we assume that for all trials in the network, treatments are coded in increasing order, so that '$k > b$' indicates that k is being compared with baseline treatment b in this trial. The model assumes that the study effects μ_i are treated as unrelated nuisance parameters. As the baseline effects can refer to different treatments in different trials, the use of informative priors on these parameters does not appear to make sense.

By setting $d_{11} = 0$, treatment 1 is taken to be the overall baseline for the network, and the treatment effects of $2, \ldots, K$, relative to treatment 1, $d_{12}, d_{13}, \ldots, d_{1K}$, emerge as our *basic* parameters. The basic parameters form a spanning tree of the network, allowing the remaining treatment contrasts, (sometimes referred to as functional parameters) to be expressed in terms of linear combinations of the basic parameters. It is noted that a slightly more general notation, which was adopted in [33], involves expressing (3) as $\delta_{ibk} \sim N(d_{bk}, \sigma^2)$, where, when assuming $d_{bk} = d_{1k} - d_{1b}$, we reach an identical parametrization. However, it is possible to express d_{bk} using any set of basic parameters that form a spanning tree of the network. In NMA, a modeling structure that allows contrasts of interest to be linked and estimated using a set of basic parameters is often referred to as the *consistency* assumption (see [32] for an extensive review).

To complete the model specification, the following vague prior distributions can be specified:

$$\begin{aligned} \sigma &\sim U(0,2), \\ d_{1k} &\overset{ind}{\sim} N(0,1000)\,,\ k = 2, \ldots, K, \\ \text{and}\ \ \mu_i &\overset{ind}{\sim} N(0,1000)\,,\ i = 1, \ldots, N. \end{aligned} \tag{4}$$

In the case of the fixed effects, the choice of a normal distribution with a large variance relative to the selected scale of the linear predictor is often a reasonable choice for a non-informative prior. Here, both the baseline effects μ_i and the treatment contrasts d_{1k} are being treated as fixed effects because the nearly improper unrelated normal priors are used and essentially preclude the shrinkage of these effects toward their grand mean. When considering the variance component σ^2, it is well known that results can be sensitive to the choice of prior distribution [9, 20]. The approach adopted in [4] follows the suggestions of [10], in that, a $U(0, 2)$ prior provides a weakly informative prior over a range of plausible values on the log relative risk scale. In analysis involving a smaller number of trials (say, five or fewer), a stronger prior such as half-Cauchy (or half-normal) with an appropriate scale parameter has been recommended [40]. Such an approach could also be adopted in cases involving a larger numbers of studies. A final possibility would be to develop an informative prior based on the large body of empirical evidence provided by Cochrane database of systematic reviews [21].

Note that in the CB models (1)–(4), the μ_i are nuisance baseline parameters. This formulation does not require pre-processed contrasts and an approximate normal likelihood, but rather merely associating a baseline treatment with each trial and parameterizing the model in terms of the contrasts.

3.2. An alternative parametrization based on a missing at random assumption and a two-way layout

Following the work on two-treatment meta-analysis in [11], an alternative way to parametrize NMA involves the use of a classical two-way linear predictor, with main effects for treatment and trial [37, 39]. As in the previous subsection and [37], the approach examined here is contrast based, with all treatment effects judged relative to a baseline. This method utilizes the same likelihood as the LA model and therefore protects against cross-level bias (i.e., it works well when trials have internal validity, but possibly limited external validity, perhaps due to differing patient populations across trials).

In the Statisticians in the Pharmaceutical Industry (PSI) working group paper [37], the two-way random effects model was applied using maximum likelihood estimation to an example where the same control treatment was available in every trial. The results were then compared with the LA Bayesian random effects model. Here, we focus on the Bayesian framework in the more general case where a common control treatment is not included in every trial. To achieve this, it is noted that all studies can in principle contain every arm, but in practice, many arms will be missing. Further, as the consistency assumption implicitly implies a missing as random assumption (missing at random) [33], a common (though possibly missing) baseline treatment can be assumed for every study [41, 42].

Returning to the two-way layout, the linear predictor is defined as

$$\eta_{ik} = s_i + t_k + \nu_{ik} \tag{5}$$

where s_i is the fixed main effect of the *ith* study, t_k is the main effect of the *kth* treatment, and ν_{ik} is a random effect associated with η_{ik}.

Following [37], restrictions are placed on one of the fixed treatment effects and the corresponding treatment-by-study random effects (note that the restriction on random effects is not required). In our case, we set $t_1 = 0$ and $\nu_{i1} = 0$ for $i = 1, \ldots, N$, where treatment 1 is again taken to be the overall baseline for the network. A multivariate normal structure is assumed for the remaining random effects:

$$(\nu_{i2}, \ldots, \nu_{iK})' \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma})\,. \tag{6}$$

To achieve a homogenous variance, analogous to (3), we can assume the $(K-1) \times (K-1)$ equicorrelated model

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}. \tag{7}$$

As with the LA random effects model, priors must be assumed for the remaining parameters. A starting point would be to assume the same vague normal priors for the fixed effects (i.e., $t_k \sim N(0, 1000)\,,\ k = 2, \ldots, K$ and $s_i \sim N(0, 1000)\,,\ i = 1, \ldots, N$) and the same uniform prior for σ. We note that because the model described in (5)–(7) focuses on an identical set of treatment contrasts to LA random effects model described in Section 3.1 (i.e., $d_{12} = t_2, \ldots, d_{1K} = t_K$), the same consistency assumptions are being made. When applying these models with non-informative priors, almost identical results would be expected. However, in the two-way model, the baseline nuisance parameters are parameterized in a different way offering much greater flexibility when forming informative priors. Specifically, in the LA model, as the set of trial baseline parameters (μ_i) are nuisance parameters with no interpretation, default 'non-informative' priors must be used. By contrast, in the two-way model, as each trial has the same parameterization $(t_1 = 0)$, an informative prior can be used for the baseline parameters s_i, allowing them to be related to the background rate associated with treatment 1.

The construction of a prior for background rate could be tackled by developing a baseline history model, a topic which is discussed in NICE technical support document 5 [43]. In this document, it is recommended that the same generalized linear modeling framework should be used to model the absolute effects of a 'standard treatment' or placebo comparator and that investigators should take care to justify their choice of data sources to inform the baseline, which could include a subset of the trials identified in the systematic review of relative effect data, cohort studies, patient registers, expert opinion, or combinations of these. Further, it is suggested that it is preferable to construct the baseline model independently from the model for relative treatment effects, in order to ensure that the latter are not affected by assumptions made about the baseline. In the context of meta-analysis of rare safety events, the prior would therefore have to be based on information outside of network to avoid double use of data sources.

The NICE document also recommended the use of a predictive distribution, rather than the fixed effect or random effects mean, to reflect the observed variation in baseline rates. When forming a prior, this idea is related to a wider literature on the discounting of historical information. In particular, the utilization of a predictive distribution from Bayesian hierarchical model to combine historical data sources, which would then subsequently form a prior for background rate (see [44] for full details). A noteworthy alternative involves adjusting the synthesis of studies identified in a systematic review for potential bias. A possible framework for bias adjustment, described in [23], provided a qualitative checklist based on assessing the rigor and relevance of each data source and then suggested the use of expert opinion to incorporate an appropriate quantitative discount. Rather than utilizing any model in the formulation of a prior, a final alternative would be direct expert elicitation using an established framework, such as the Sheffield elicitation framework [45–47]. In this case, it would be important to emphasize to the experts that their prior should be using information external to the network.

As we do not have an appropriate systematic review to build a baseline prior for the NSAID case study, in the subsequent analysis presented in Section 5, we shall not directly apply any of these approaches. The review presented in this section serves as a guide to possible methods. However, it is acknowledged that practical experience in this area is somewhat limited.

3.3. Multivariate random effects with borrowing strength across outcomes

Suppose that in each trial, we now have *multiple* binary outcomes indexed by ℓ, where $\ell = 1, \ldots, L$. Thus, our Poisson event rate λ_{ik} in (1) now becomes $\lambda_{ik\ell}$, with corresponding log-event rate $\eta_{ik\ell} \equiv \log \lambda_{ik\ell}$ (note that alternative sampling models could be used for different types of outcomes, but the following principles still apply). Under multiple outcomes, $\ell = 1, \ldots, L$, meaning that our Poisson likelihood model can be rewritten as

$$Y_{ik\ell} \sim \text{Poisson}(E_{ik\ell}\lambda_{ik\ell}),$$

where i and k index studies and treatments as before, but now, ℓ indexes the L outcomes.

Up until now, our models have all been contrast based, in the sense that they assume a baseline study effect μ_i augmented by a treatment effect d_{1k} (for fixed effect models) or δ_{ibk} (for random effect models) that is defined to be identically 0 when $k = b$. While extremely common in the literature and existing software, the approach is not without its drawbacks. First, the baseline effects are hard to interpret because the baseline treatment b often changes across trials. The method is also asymmetric in its handling of the treatments; note for instance from (2) that $Var(\eta_{ib}) < Var(\eta_{ik})$ for all $k > b$, even though there is no substantive reason to suspect such a relationship.

As such, several authors [7, 41] have proposed an *arm-based* (AB) parametrization, in which the effect of each treatment is modeled directly, rather than relative to a particular baseline. In this framework, we replace model (2) for η_{ik} with

$$\eta_{ik\ell} = \log(\lambda_{ik\ell}) = \mu_{k\ell} + \nu_{ik\ell}\,, \tag{8}$$

where $\mu_{k\ell}$ is the fixed mean effect of treatment k with respect to outcome ℓ, and $\nu_{ik\ell}$ is the study-specific random effect, again for outcome ℓ. If independence across outcomes (but not treatments) is reasonable, we might assume the $\boldsymbol{\nu}_{i\ell} = (\nu_{i1\ell}, \ldots, \nu_{iK\ell})^T$ vectors to be independently distributed as $MVN\left(\mathbf{0}, \boldsymbol{\Lambda}_\ell^{Trt}\right)$, where $\boldsymbol{\Lambda}_\ell^{Trt}$ is a $K \times K$ unstructured covariance matrix capturing relations of random effects between treatments. In what follows, we refer to this model as ABRE1. Alternatively, we might assume dependence of these random effects across outcomes ℓ but independence across treatments. This would lead us to redefine the random effect vectors as $\boldsymbol{\nu}_{ik} = (\nu_{ik1}, \ldots, \nu_{ikL})^T$, and assume them to be independently distributed *a priori* as $MVN\left(\mathbf{0}, \boldsymbol{\Lambda}_k^{Out}\right)$ where $\boldsymbol{\Lambda}_k^{Out}$ is an $L \times L$ covariance matrix capturing relations between outcomes for the k^{th} treatment. We refer to this model as ABRE2. Note that we may assume $\boldsymbol{\Lambda}_\ell^{Trt} = \boldsymbol{\Lambda}^{Trt}$ for all ℓ or $\boldsymbol{\Lambda}_k^{Out} = \boldsymbol{\Lambda}^{Out}$ for all k when it is reasonable to do so.

Finally, in order to capture *both* correlations across treatments and outcomes, we would need to modify (8) to

$$\eta_{ik\ell} = \log(\lambda_{ik\ell}) = \mu_{k\ell} + \nu_{ik} + \omega_{i\ell}\,, \tag{9}$$

where $\boldsymbol{\nu}_k \equiv (\nu_{i1}, \ldots, \nu_{iK})^T \sim MVN(\mathbf{0}, \mathbf{D}^{Trt})$, $\boldsymbol{\omega}_\ell \equiv (\omega_{i1}, \ldots, \omega_{iL})^T \sim MVN(\mathbf{0}, \mathbf{D}^{Out})$, and the two random effect specifications are mutually independent. Here, $\mathbf{D}^{Trt}$ and $\mathbf{D}^{Out}$ are $K \times K$ and $L \times L$ unstructured covariance matrices. We refer to model (9) as ABRE3. Note the lack of any triply-subscripted effects, in order to control their total number and preserve identifiability.

Regarding prior distributions, following common modern Bayesian practice, we recommend starting with 'default', typically minimally informative priors. For example, we typically begin by assigning independent vague normal distributions to the $\mu_{k\ell}$, namely $N(0, 100^2)$ priors. Similarly, we often initially assume that all inverse covariance matrices follow vague Wishart distributions with degrees of freedom γ and a mean of $\gamma \boldsymbol{\Omega}^{-1}$, where we select $\boldsymbol{\Omega}$ and γ to be the identity matrix and $\dim(\boldsymbol{\Omega})$, respectively, yielding extremely vague but still proper priors. To check prior sensitivity, we often switch to moderately informative Wishart priors by changing our $\boldsymbol{\Omega}$ and γ selections, for instance, by setting the off-diagonal elements of $\mathbf{\Omega}$ equal to 0.5 or -0.5 instead of zero, and the degrees of freedom γ to 40. The mean parameters $\mu_{k\ell}$ can also be set to more sensible, non-zero values appropriate in the given applied context; in what follows, we simply keep these priors noninformative.

By dropping the dependence on any particular baseline treatment, the AB models *seem* to implicitly assume that data will be available on all treatments for all studies. Of course, this will almost never be the case in practice, but AB models can still be easily specified and fit under a *missing data* framework, wherein we use our full Bayesian model to impute the values of any missing data as part of our MCMC implementation. Such an approach is tantamount to assuming that any missing data are missing at random [48, 49], but nonrandom missingness could be accommodated if we were also willing to specify some sort of selection model. In any case, this AB approach does implicitly assume consistency between direct and indirect evidence, as well as comparability of study populations across trials (and not merely comparability of contrasts with each trial's baseline). Fortunately, the former assumption can be checked because numerical discrepancies between any imputed values and corresponding estimates obtained indirectly would suggest lurking inconsistency. The definition and remedying of inconsistency in AB methods are areas of ongoing research and subjects to which we return in Section 6.

3.4. Models for multiple outcomes and finding the best treatment

The primary goal of MTCs is often to identify the 'best' treatment. However, this determination can be difficult when we are measuring *multiple* outcomes on each individual. For instance, an MTC might measure safety in two different ways, by measuring both the presence or the absence of a particular unpleasant side effect, and also whether or not an individual was forced to discontinue the study drug for any reason. Certain drugs might do well on the former measure but poorly on the latter, and vice versa for other drugs. In this setting, the hierarchical Bayesian machinery is well suited to making an overall decision based on the totality of evidence.

Building upon the multiple outcome setting described in Section 3.3, a natural Bayesian summary is the posterior probability of being best under each outcome,

$$\begin{aligned} &Pr\{\text{k is the best treatment for outcome } \ell \mid \mathbf{y}_\ell\} \\ &\quad = Pr\{\text{rank}(d_{1k\ell}) = 1 | \mathbf{y}_\ell\}\,, \end{aligned} \tag{10}$$

a quantity we denote by 'Best1'. Here, we adopt the convention that low ranks are good because the outcome has a negative interpretation (adverse events). Similarly, we might calculate the probability of being the first *or* second best treatment, say denoted by 'Best12', by replacing the right-hand side of equation (10) with $Pr\{\text{rank}(d_{1k\ell}) = 1 \text{ or } 2 \mid \mathbf{y}_\ell\}$, where again small ranks indicate best treatments.

Next, to integrate these univariate probabilities over all the outcomes and obtain one omnibus measure of 'best', we can use an overall, weighted score, defined as

$$S_k = \sum_\ell w_\ell \, d_{1k\ell}\,, \tag{11}$$

where w_ℓ is the weight for outcome ℓ, and $\sum_\ell w_\ell = 1$. Writing $\mathbf{y} \equiv \{\mathbf{y}_\ell,\ \ell = 1, \ldots, L\}$, the weighted scores S_k can be used to obtain overall Best1 probabilities as

$$Pr\{\text{k is the best treatment overall} \mid \mathbf{y}\} = Pr\{\text{rank}(S_k) = 1 | \mathbf{y}\}\,,$$

and similarly for Best12. Note that the weights w_ℓ are *not* estimated from the data but chosen by clinicians or other team leaders based on their relative preferences among the safety outcomes.

In practice, we recommend investigating several different weights to see if the decision regarding the best treatment is sensitive to the weighting scheme. In addition, we caution that, while a single drug may emerge with a fairly large Best1 or Best 12 probability (say, 0.7), there is often substantial uncertainty associated with these estimates. Thus, a drug that appears to have a very high chance of best may not be *significantly* better than several others in the MTC. As such, we further recommend looking at side-by-side boxplots of the $d_{1k\ell}$ and S_k posterior distributions, in order to judge their overlap. Finally, in settings where placebo is *known with certainty* to be the safest medication, it should be removed from the ranking procedure (though not of course from the data analysis itself).

4. CONSIDERATIONS WHEN REPORTING A BAYESIAN SAFETY META-ANALYSIS

As described earlier in this paper, there are a number of advantages that the Bayesian approach offers in the context of meta-analysis applied to drug safety. In this section, we propose a checklist that will be useful in ensuring standardization of the reporting process of Bayesian meta-analyses, particularly in the regulatory setting. This checklist is important for a number of reasons. First, there is limited consensus regarding the reporting of Bayesian meta-analysis, in part because within medical product development, it is recently emerging among statisticians and is largely unknown outside the statistical community. Second, Bayesian methods offer important advantages in flexibility and complexity relative to frequentist analysis and therefore require certain aspects be reported and discussed. Third, utilization of such a list could improve the scientific rigor and ensure transparency, thereby enabling replication and verification by other investigators and stakeholders. Fourth, standardized reporting also can promote a consistent approach so that all stakeholders (regulators, sponsors, investigators, the public, etc) can understand the evidence synthesis more fully. Fifth, the checklist can provide guidance to reviewers and readers to better understand such analyses. Finally, given that the terminology tends to be different between a Bayesian and a traditional analysis, standardized reporting promotes the use of similar terminology and provides clear distinctions whenever a Bayesian meta-analysis is reported.

There are guidelines available for the implementation of traditional meta-analysis, such as the PRISMA statement [50, 51], as well as the general meta-analysis guidelines for the presentation of meta-analysis results (see [52], Chapter 11). The concepts considered in the process of a systematic review (such as how to formulate a study question, conduct literature searches, etc.) are completely relevant for the conduct of any meta-analysis, regardless of whether the analytical approach is Bayesian or frequentist (see [52], Chapters 5–8).

A general Bayesian checklist, together with a series of examples, has also been produced in the context of health technology assessment [10]. In addition, a list for the review of evidence syntheses of treatment efficacy used in decision making, based on either efficacy or cost-effectiveness, has been recently provided in a NICE technical support document [53]. Many of the principles are relevant to a meta-analysis for safety data. However, a Bayesian meta-analysis of a safety outcome has some unique considerations that may differ from a synthesis of efficacy data.

Along with the work on meta-analysis and evidence synthesis, there is also a limited parallel literature on the Bayesian reporting of clinical trials and observational studies. For example, Lang and Secic [54] provided the following fundamental elements for Bayesian reporting: a) report the pre-trial probabilities, and specify how they were determined; b) report the post-trial probabilities and the corresponding intervals; and c) interpret the post-trial probabilities. Similar advice is provided in the *Annals of Internal Medicine*'s instructions to authors for manuscripts using Bayesian methods to conduct data analyses. Further, a list of seven items that 23 international experts believe to be most important when reporting a Bayesian analysis for scientific publications has been generated [55]. The set of items were as follows: describing the prior distribution, through specification, justification, and sensitivity analysis; presenting the analysis in terms of the statistical model and analytic technique; and presenting the results using a measure of central tendency and variance.

Based on the preceding literature review, we combine the ideas on conduct and reporting of meta-analysis with more general guidance on Bayesian reporting to produce the checklist for the conduct and reporting of a Bayesian meta-analysis shown in Table I. Many of the items in the checklist, including 1–3, 5, 6, 9, 11, and 14, overlap with and should be considered complementary to other established clinical trial reporting guidelines, such as the CONSORT statement [56], as well as general meta-analysis reporting guidelines. By contrast, items 4, 7, 8, 10, 12, and 13 reflect unique aspects of Bayesian meta-analysis reporting in the context of safety data.

Table I includes four main sections: Introduction, Methods, Results, and Interpretation. Each main section includes various items relevant to that section. The user of the table should evaluate each item and may wish to utilize the last two columns (left blank in our template) to confirm whether or not each item has been addressed and to add any relevant comments. The Introduction section includes items that are specific to the intervention being analyzed along with the meta-analysis objective(s), while the Methods section focuses on aspects associated with the planned modeling and analyses to be conducted. The Results section relates to specific summaries of the data that should be provided. Finally, the Interpretation section enables decision makers utilizing the meta-analyses to appropriately evaluate the results.

5. AN ILLUSTRATION OF THE DEVELOPMENT AND REPORTING OF A BAYESIAN NETWORK META-ANALYSIS

5.1. Background to non-steroidal anti-inflammatory drugs example

Non-steroidal anti-inflammatory drugs are drugs that provide analgesic and antipyretic effects and, in higher doses, anti-inflammatory effects. The most prominent members of this group of commonly used drugs are aspirin, ibuprofen, and naproxen, all of which are available over the counter in most countries. NSAIDs inhibit cyclooxygenase (COX)-mediated production of prostaglandins. Concerns about the cardiovascular safety of NSAIDs arose since the withdrawal of rofecoxib [57], a selective COX-2 inhibitor, from the market in 2004 after the results of a randomized placebo-controlled trial showed an increased risk of cardiovascular events associated with the drug [58]. This finding was confirmed in other trials and a cumulative meta-analysis [2]. However, observational studies have provided conflicting data on the association of rofecoxib with cardiovascular risk [59–64]. In an effort to provide a more synchronized and comprehensive evaluation of the cardiovascular safety of NSAIDs, recently, an NMA was conducted to integrate all available randomized evidence [4]. Additionally, systematic reviews of observational studies on cardiac risk of NSAIDs have also been conducted [57,65,66].

The recent systematic review [4] of large-scale RCTs comparing any NSAID with other NSAIDs, paracetamol (acetaminophen), or placebo for any medical condition uncovered 31 RCTs evaluating seven different NSAIDs. Multiple cardiac outcomes were assessed in the NMA, with fatal or nonfatal myocardial infarction (MI) as the primary outcome. A Bayesian random effects model was used, with trials having zero events in both groups excluded from the analysis. Detailed information of the included studies can be found in [4]. For illustration, in this case study, we will focus solely on the MI and stroke outcomes. Thirty studies had MI and stroke collected. The patient years and number of MI events by randomized intervention are provided in Table II.

Figure 1 exhibits the network of seven NSAIDs with placebo for both the MI and the stroke outcomes. The size of each node denotes the total number of trials studying the drug. The thickness of each edge represents the total patient years, with the number of studies for the relation on each edge. As can be seen in Table II, only four NSAIDs (naproxen, celecoxib, rofecoxib, and lumiracoxib) are connected to placebo; two thirds of the trials do not include placebo as their control arm. For the analysis, each drug is assigned an index from 1 to 8 for placebo to lumiracoxib in a clockwise direction around Figure 1.

5.2. Results

We first applied a number of the models and parameterizations described in Subsections 3.1–3.3 to the stroke and MI data separately. Specifically, we considered an AB multivariate random effects model, the LA fixed effects model, the LA random effects analysis utilized in [4], and the corresponding random effects model in a two-way layout utilizing both a standard prior and a multivariate prior for the fixed effects. In addition, a naive pooling of the data was investigated for comparison purposes. When fitting these models, and in all subsequent models presented in this section, WinBUGS 1.4.3 [67] was used to draw 50,000 MCMC samples from each of two parallel sampling chains following a 50,000-iteration burn-in period . Because smaller datasets or models that avoid imputing missing data (i.e., LA models) converge relatively fast, chains this long may not be required in all cases. However, we need enough samples to ensure MCMC convergence when the data are large and complex or when missing data are imputed. To check convergence, we used several standard diagnostics, including lag 1 sample autocorrelations and visual inspection of sample trace plots.

The results depicted in forest plots of the risk ratios relative to placebo (Figures 2 and 3) show that the LA fixed effects model and all random effects models that associated a separate baseline fixed effect with each trial provide similar results for the stroke and MI data. By contrast, results from the AB model often fall between those of the other models and those based on naive

Table I. Checklist for Bayesian safety meta-analysis.

Item number	Key elements for reporting Bayesian meta-analysis	Description	Item completed?	Comments
	Introduction			
1	Intervention and population of interest	The intervention to be investigated with respect to the relevant population		
2	Aim of study	Research questions and objectives of the meta-analysis		
	Methods			
3	Study design	Special attention should be given to the similarity of studies in order to justify any assumptions of exchangeability.		
4	Prospective Bayesian analysis?	Whether the prior was constructed before the data collection and analysis		
5	Outcome measure	The true underlying parameters of interest for analysis and reporting		
6	Statistical model	Likelihood function; whether posterior was derived by simulation or analytical methods; rate of missingness and imputation of any missing covariates; structure of levels of model; choice of fixed effect versus random effects models and rationale for choice		
7	Prior distribution	Choice of prior distributions (and hyperprior distributions if hierarchical modeling is used) and rationale for choices; alternative priors for the purpose of sensitivity analysis should be explicitly specified.		
8	Computation/software	Computational methods for generating posterior inferences; software used and how it is validated if not commercially available; if MCMC is used, how convergence is checked.		
9	Planned analyses for model checking, prior to posterior sensitivity, and convergence diagnostics	The approaches used to check model fit and to carry out any sensitivity analyses		
	Results			
10	Describe posterior distribution of parameters and other quantities of interest.	Summaries (numerical and/or graphical) of the posterior distribution of study specific and overall model parameters and other quantities of interest		
11	Results for model checking and convergence diagnostics	Findings of these analyses and implications for study results		
	Interpretation			
12	Bayesian interpretation	Bayesian interpretation of the results with respect to central tendency, standard deviation, or credible interval of the parameters of interest based on the posterior distribution or posterior predictive probabilities of certain hypothesis statements		
13	Impact of prior to posterior sensitivity	The results of any alternative priors and/or addressing the impact of choice of priors on the conclusions		
14	Possible limitations of the analysis	Include an honest appraisal of the strengths and possible weaknesses of the analysis		

Table II. Data for meta-analysis of non-steroidal anti-inflammatory drug trials [4]: patient years (Pt-Yrs), and event counts (*n*) using the format 'MI; stroke'.

	Placebo		Naproxen		Ibuprofen		Diclofenac		Celecoxib		Etoricoxib		Rofecoxib		Lumiracoxib	
Trial	Pt-Yrs	*n*	Pt-Yrs	*n*	Pt-Yrs	*n*	Pt-Yrs	*n*	Pt-Yrs	*n*	Pt-Yrs	*n*	Pt-Yrs	*n*	Pt-Yrs	*n*
ADAPT	1982	13; 7	1332	13; 10					1346	8; 7						
Aisen, 2003	115	1; 1	124	0; 5									126	3; 4		
Geusens, 2004	111	0; 1	118	0; 0											234	2; 3
APC	1558	3; 3							3124	18; 8						
IQ5-97-02-001	120	0; 3							285	2; 7						
PreSAP	1570	4; 7							2331	9; 9						
Lehmann, 2005	98	1; 0							99	0; 0					200	0; 1
APPROVe	5711	18; 9											5658	34; 19		
Reines, 2004	293	4; 5											273	2; 1		
Thal, 2005	1820	13; 15											1599	22; 7		
VICTOR	986	1; 3											928	6; 3		
ViP	1102	5; 3											1099	6; 2		
A3191152			130	0; 0					131	0; 0						
SUCCESS-1 (USA/Canada)			165	1; 2					353	4; 2						
ADVANTAGE			526	1; 6									528	5; 1		
VIGOR			2008	4; 9									2007	20; 11		
TARGET (0117)			4156	7; 13											4197	15; 17
CLASS (N49-98-02-035)					1123	9; 6			1184	9; 2						
TARGET (2332)					3709	5; 9									3795	5; 8
CAESAR							432	5; 5	415	4; 1						
CLASS (N49-98-02-102)							1081	5; 6	1136	10; 2						
Emery, 1999							125	0; 0	133	1; 0						
SUCCESS-1 (World)							745	0; 4	1472	6; 6						
EDGE							2607	11; 6			2789	19; 4				
EDGE II							3251	25; 12			3266	14; 9				
MEDAL							19,103	88; 42			19,970	84; 48				
Cannon, 2000							256	1; 1					494	2; 0		
Saag, 2000							219	1; 1					443	2; 1		
Fleischmann, 2003									99	0; 0					207	2; 0
Tannenbaum, 2004									110	0; 0					225	1; 0
Overall	15,466	63; 57	8559	26; 45	4832	14; 15	27,819 0	136; 77	12,218	71; 44	26,025	117; 61	13,155	102; 49	8858	25; 29

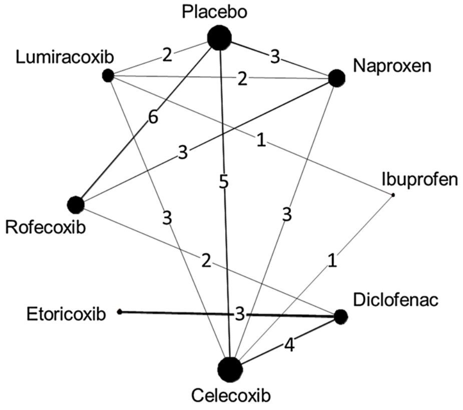

Figure 1. Network graph of non-steroidal anti-inflammatory drugs in safety meta-analysis.

pooling. This is unsurprising as the AB model allows for some borrowing of strength across control groups, although clearly not to the extent of naive pooling.

To further investigate the data, we fit three AB models with different correlation structures, introduced in Subsection 3.3, and compare the results to those from the LA random effects model. The upper part of Table III shows model-specific DIC scores, which we recall represent a compromise between model fit ($\overline{D}$) and effective complexity (p_D). The LA random effects model fits the data best with the smallest $\overline{D}$, although it gives the largest p_D value. ABRE1 assumes independence between the two outcomes, as in the LA random effects model, but utilizes a missing data imputation framework. ABRE2 adopts correlation across outcomes. Smaller p_D are observed for these two models than in the LA random effects model, although their DICs are not appreciably different (i.e., the difference between them is smaller than 5). However, ABRE3 (which permits correlation across both outcomes and treatments) does give an appreciably smaller DIC, thanks to a p_D 11 units smaller than that of the LA random effects model, apparently the result of shrinkage across the two sets of random effects.

The lower part of Table III shows rate ratios (RRs) of the seven NSAIDs to placebo with associated 95% credible intervals. Here, an RR smaller than 1 indicates a safer drug than placebo in terms of MI or stroke occurrence. For the MI outcome, the three AB models agree on the safety ordering of the drugs, but this ordering differs from that of the LA random effects model; several RRs (for ibuprofen, diclofenac, etoricoxib, and lumiracoxib) are 'flipped' to the other side of 1. Compared to the LA random effects model, the RRs for ibuprofen and lumiracoxib under AB models decrease about 50%. We hasten to add that only rofecoxib's RR is significantly greater than 0, and even this significance is quite marginal for the ABRE models. Regarding the stroke outcome, none of drugs is safer than placebo under the LA random effects model with several fairly high RRs (ibuprofen, diclofenac, etoricoxib, and lumiracoxib), while all drugs except naproxen are safer than placebo under ABRE1 and ABRE2. Still, it is important not to overinterpret this finding, as none of the 95% Bayesian credible intervals in this part of the table exclude 1.0. Some RRs in ABRE3 (ibuprofen, diclofenac, and lumiracoxib) are 'flipped' from ABRE1 and ABRE2, although they remain close to 1.

Turning to overall rankings of the NSAIDs, we calculate Best12 probabilities for each drug by incorporating the weighted score (11). We set the MI outcome to be $\ell = 1$ and investigate three different weights: $w_1 = 0.5, 0.8,$ and 0.2. That is, we consider equal weight on both outcomes with $w_1 = 0.5$, more weight on MI than stroke under $w_1 = 0.8$, and vice versa under $w_1 = 0.2$ (because here, $w_2 = 1 - w_1$). Because ABRE3 emerges as the DIC-best model among all AB models, Table IV reports only Best12 probabilities from the LA random effects and ABRE3 models. In the former case, placebo is the overall winner across all weighting schemes, probably the result of the consistently high stroke RRs for all non-placebo drugs in this model. By contrast, the ABRE3 winner varies with different weights: Naproxen is safest under $w_1 = 0.8$, but its slightly higher stroke risk causes etoricoxib to emerge as best when $w_1 = 0.5$ or 0.2 (i.e., we place greater emphasis on stroke safety). Still, the Best12 probabilities for the drugs in second and third places are close to those for winners, the result of the high degree of uncertainty in these data. Here, we believe that the discrepancy between the winners under the LA and ABRE models occurs because of the constraint imposed by CB models like LA. That is, the LA model assumes that every study has its own (fixed) baseline effect regardless of whether it is placebo or active drug, whereas AB models estimate baseline (in this case, always placebo) effects incorporating all available observed and missing information. Here, because only $1/3$ of our studies have placebo as their baseline treatment, the estimated placebo effect in the LA model could be somewhat different from that of the AB model, and this might lead to different decisions.

The LA model gives $\hat{\sigma} = 0.34$ (0.01, 0.89) and 0.27 (0.02, 0.73) for the MI and stroke outcomes, respectively, while ABRE2 gives 0.50 (0.35, 0.70) and 0.87 (0.62, 0.87) with 0.5 (0.07, 0.79) correlation between these two outcomes. ABRE3 leads to a bit larger variability because it has two sources of uncertainty, with 0.31 (-0.36, 0.76) between-outcome correlation. To check sensitivity of our findings to the prior specification, we switch to moderately informative Wishart priors on the random effect covariance matrices for ABRE2 and ABRE3, as we described in Section 3.3. We force the prior mean of diagonal elements of the covariance matrix to be 1. For ABRE2, these alternative priors do not yield better DIC scores nor affect the mean parameters $\mu_{k\ell}$ very much, even when $\boldsymbol{\Omega}$ induces 0.5 between-outcome correlation. The estimated variabilities increased to 0.8 for both outcomes, yielding slightly wider credible intervals of $\mu_{k\ell}$. For ABRE3, we checked three different priors: adding the moderately informative prior on $\mathbf{D}^{Out}$ but not $\mathbf{D}^{Trt}$, on $\mathbf{D}^{Trt}$ but not $\mathbf{D}^{Out}$, and on both of these matrices, in all cases forcing the corresponding covariance matrices to have variance 1 and correlation 0.5. As we observed in ABRE2, all three cases do not produce smaller DIC than under the noninformative prior. Overall, the estimated $\mu_{k\ell}$ is similar to that estimated with the noninformative prior, although the Best12 probabilities are changed a bit without switching the winner, except for the first case (moderately informative prior on $\mathbf{D}^{Out}$ alone). In this case, placebo now emerges as the safest drug, with Best12 probability 0.4 when $w = 0.5$. Again, no narrower credible interval width is observed because the estimated variability is somewhat larger than that under noninformative priors.

6. DISCUSSION

The case study presented in this paper reconsidered a previously published NMA examining cardiovascular outcomes associated with the use of NSAIDs. In our analysis, we focused on the impact of different model choices, the interplay between prior and parametrization, and extensions based on the analysis of

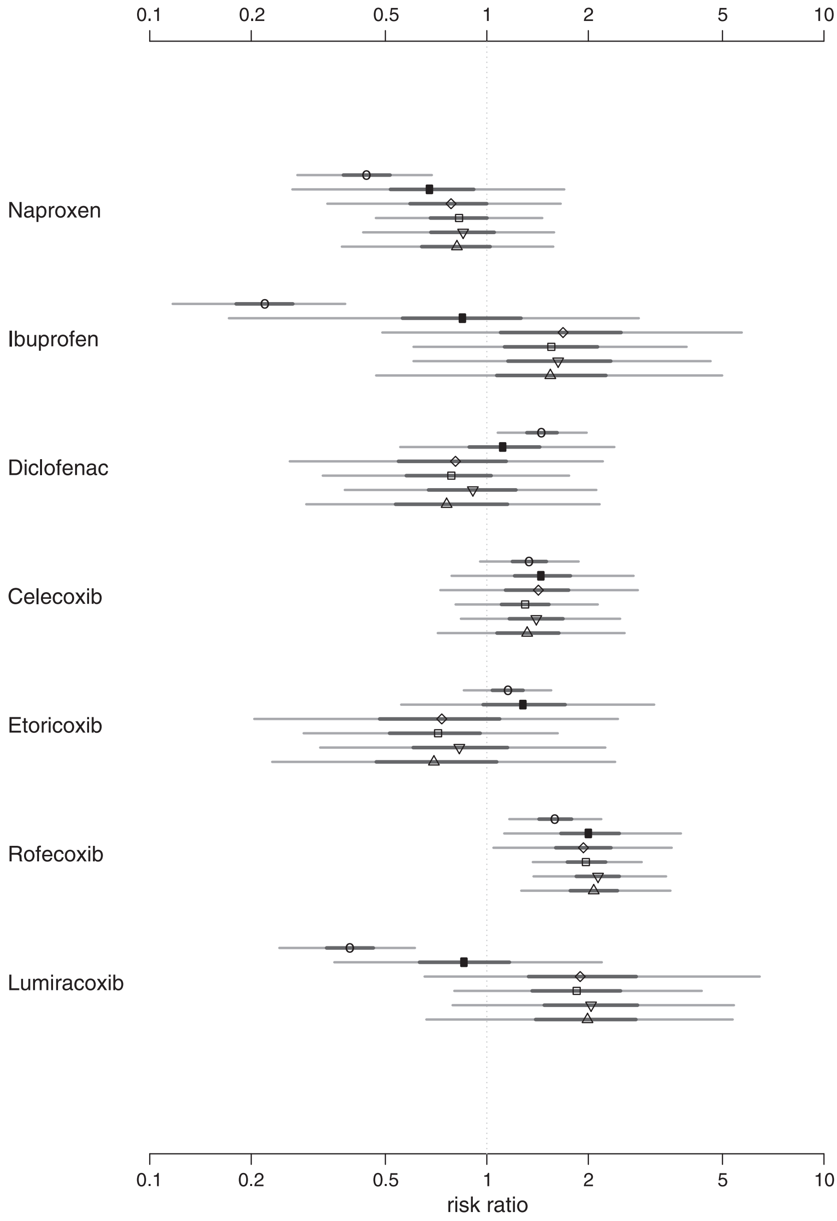

Figure 2. Analysis of the myocardial infarction data provided in the non-steroidal anti-inflammatory drug network. Six model fits are displayed, with the posterior 95% intervals and 50% intervals (overlaid), and a symbol marking the median. The models are naive pooling (circle), arm based (solid circle), Lu and Ades (LA) random effects analysis of Trelle *et al.* (diamond), LA fixed effects (square), two-way layout with standard prior (triangle point down), and two-way layout with multivariate prior (triangle point up).

multiple outcomes. The analysis showed that the outcomes can be particularly sensitive to the choice of model and outcome weighting scheme, emphasizing the need for model sensitivity analysis along with transparency regarding assumptions and a fair appraisal of the potential limitations when reporting results.

When presenting the analysis, we did not focus on examining the assumption of consistency that has to be made when fitting the standard fixed and random effects models, either with the contrast based or with the AB alternatives. However, in the previously published analysis [4], a careful examination of consistency is provided. More generally, a number of recently published guidelines provide a thorough overview of this area [7, 8].

One of the potential weaknesses of our NSAIDs NMA is that it was unable to account for different dose levels. This problem is generally difficult to deal with in NMA of summary data. A possible solution is examined in the recent work presented in [68], where models are developed that combine NMA with dose–response modeling. A further limitation of the NSAID case study was the assumption of a constant treatment effect across the time a patient is exposed to a treatment, which is necessary when working with summary exposure data and a Poisson sampling model. To fully examine this assumption, individual level patient data (IPD) meta-analysis is required. Such an approach permits time-to-event analyses to examine long-term outcomes and assess the possibility of non-proportional hazards. In addition, IPD meta-analysis offers a number of other advantages, including the use of common definitions, coding and cutpoints to produce consistent analyses across multiple studies, the

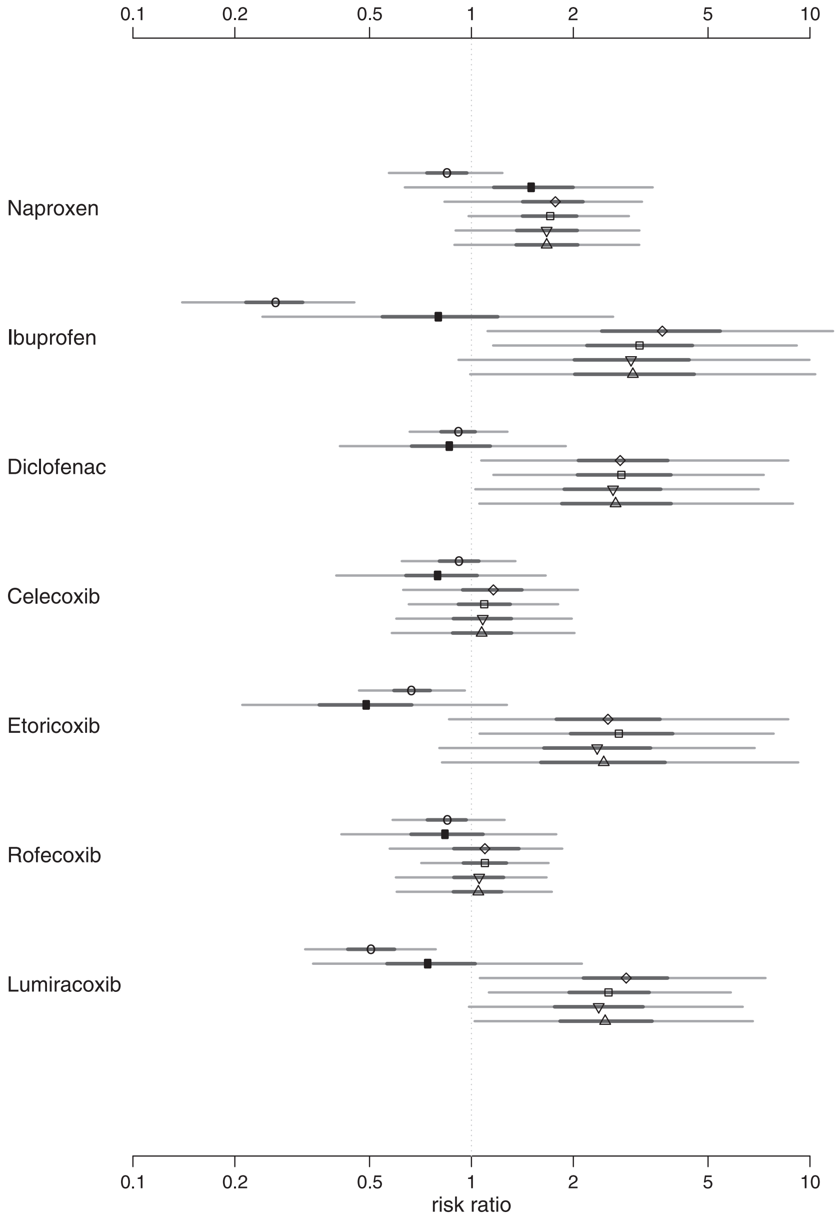

Figure 3. Analysis of the stroke data provided in the non-steroidal anti-inflammatory drug network. Six model fits are displayed, with the posterior 95% intervals and 50% intervals (overlaid), and a symbol marking the median. The models are naive pooling (circle), arm based (solid circle), Lu and Ades (LA) random effects analysis of Trelle *et al.* (diamond), LA fixed effects (square), two-way layout with standard prior (triangle point down), and two-way layout with multivariate prior (triangle point up).

exploration of heterogeneity at the patient level and subgroup analyses of patient level data, and the investigation of additional hypotheses (particularly related to individual patient characteristics) where the data would be lacking in published results. For these reasons, IPD meta-analysis is regarded as the gold standard and, when feasible, should be considered [69, 70].

As an example of Bayesian IPD meta-analysis, we highlight a recent meta-analysis involving data from multiple companies that was conducted to investigate the cancer risk in three tumor necrosis factor (TNF) inhibitors, per EMA request [71]. Seventy-four RCTs that compared TNF inhibitors across multiple indications for at least 4-week durations were included ($n = 22,904$). Bayesian hierarchical piecewise exponential survival models were used by dividing time into multiple intervals, with constant hazard within each interval, allowing for relaxing the proportional hazards assumption. Class effects and drug-specific effects among three anti-TNF agents were assessed. The Bayesian approach provided a very powerful and cohesive framework to model the uncertainty of all parameters in this IPD meta-analysis setting. The analysis not only took into account the patient-level covariates including time-dependent covariates and between-study heterogeneity but also adaptively modulated the extremes in the rare adverse event setting, so the inferences would be more robust and reliable.

Another limitation of our data analysis is that all our models fit the data under the consistency assumption. Checking consistency is technically challenging and an area of ongoing research activity, especially for AB models. Several authors have suggested

Table III. Deviance information criterion and rate ratios of non-steroidal anti-inflammatory drugs to placebo with 95% credible interval in the parenthesis under Lu and Ades random effects and three AB models.

	LA random effects	ABRE1	ABRE2	ABRE3
Dbar	119.6	125.5	129.4	122.3
p_D	77.2	73.1	65.2	66.3
DIC	196.8	198.6	194.6	188.6
Rate ratio (95% credible interval)				
MI				
Naproxen	0.83 (0.38, 1.67)	0.69 (0.25, 1.72)	0.73 (0.34, 1.53)	0.73 (0.30, 1.67)
Ibuprofen	1.70 (0.53, 5.90)	0.88 (0.21, 3.98)	0.77 (0.28, 2.08)	0.88 (0.22, 3.53)
Diclofenac	0.85 (0.30, 2.36)	1.15 (0.52, 2.47)	1.13 (0.60, 2.14)	1.14 (0.51, 2.43)
Celecoxib	1.41 (0.74, 2.80)	1.45 (0.76, 2.82)	1.33 (0.73, 2.41)	1.31 (0.67, 2.46)
Etoricoxib	0.78 (0.24, 2.56)	1.26 (0.49, 3.34)	1.15 (0.54, 2.50)	1.02 (0.38, 2.76)
Rofecoxib	2.11 (1.25, 3.59)	1.97 (1.04, 3.83)	1.83 (1.00, 3.34)	1.84 (0.97, 3.46)
Lumiracoxib	2.11 (0.73, 6.35)	0.82 (0.30, 2.16)	0.71 (0.32, 1.59)	0.88 (0.32, 2.35)
Stroke				
Naproxen	1.70 (0.90, 3.27)	1.55 (0.63, 3.61)	1.51 (0.69, 3.26)	1.43 (0.62, 3.20)
Ibuprofen	3.22 (0.97, 11.84)	0.82 (0.21, 3.11)	0.79 (0.25, 2.47)	1.18 (0.28, 5.12)
Diclofenac	2.72 (0.95, 8.29)	0.92 (0.42, 2.08)	0.82 (0.39, 1.73)	1.02 (0.43, 2.48)
Celecoxib	1.09 (0.59, 2.05)	0.84 (0.39, 1.71)	0.79 (0.38, 1.57)	0.84 (0.41, 1.65)
Etoricoxib	2.52 (0.74, 8.63)	0.54 (0.17, 1.60)	0.51 (0.19, 1.31)	0.91 (0.30, 2.89)
Rofecoxib	1.05 (0.60, 1.74)	0.88 (0.39, 1.91)	0.82 (0.39, 1.69)	0.93 (0.47, 1.82)
Lumiracoxib	2.64 (1.00, 7.38)	0.79 (0.27, 2.26)	0.77 (0.31, 1.88)	1.06 (0.38, 2.97)

LA, Lu and Ades; DIC, deviance information criterion; MI, myocardial infarction.

Table IV. Decision making of non-steroidal anti-inflammatory drugs data; Best12 probability with weighted score under the Lu and Ades random effects and ABRE3 models.

	$w_1 = 0.5$	$w_1 = 0.8$	$w_1 = 0.2$
LA random effects			
Placebo	**0.885** (0.32)	**0.673** (0.47)	**0.880** (0.32)
Naproxen	0.386 (0.49)	0.608 (0.49)	0.107 (0.31)
Ibuprofen	0.020 (0.14)	0.032 (0.18)	0.018 (0.13)
Diclofenac	0.070 (0.26)	0.210 (0.41)	0.029 (0.17)
Celecoxib	0.389 (0.49)	0.105 (0.31)	0.534 (0.50)
Etoricoxib	0.170 (0.38)	0.355 (0.48)	0.087 (0.28)
Rofecoxib	0.072 (0.26)	0.006 (0.08)	0.335 (0.47)
Lumiracoxib	0.007 (0.09)	0.012 (0.11)	0.010 (0.10)
ABRE3			
Placebo	0.291 (0.45)	0.244 (0.43)	0.251 (0.43)
Naproxen	0.234 (0.42)	**0.449** (0.50)	0.096 (0.29)
Ibuprofen	0.353 (0.48)	0.397 (0.49)	0.296 (0.46)
Diclofenac	0.188 (0.39)	0.138 (0.35)	0.208 (0.41)
Celecoxib	0.168 (0.37)	0.062 (0.24)	0.327 (0.47)
Etoricoxib	**0.372** (0.48)	0.310 (0.46)	**0.389** (0.49)
Rofecoxib	0.031 (0.17)	0.007 (0.08)	0.134 (0.34)
Lumiracoxib	0.363 (0.48)	0.393 (0.49)	0.300 (0.46)

The ‘best’ treatment is in bold.
LA, Lu and Ades.

various methods to define and measure inconsistency. Both Lu and Ades [32] and Dias *et al.* [43] proposed addition of new random effects called *w-factors*, defined by the number of independent ‘loops of evidence’ in the NMA network graph. Evidence that these factors are significantly different from zero implies inconsistency. An alternative approach, described in [72], introduces *node-splitting* method, which enables a test of consistency through the posterior distribution of an additional parameter measuring discrepancy between direct and indirect information. In a similar vein, Presanis *et al.* [73] identify a 'separator node' in the model's directed acyclic graph (DAG) and test whether the two resulting portions of the DAG result in the same inference about the separator node. Recently, Piepho *et al.* [39] define inconsistency differently, namely as an interaction between trial types and treatments, with such interactions estimated using a two-way linear mixed model. However, a common problem of all these methods is that such inconsistency-related parameters (i.e., *w*-factors and interactions) are difficult to estimate when data are sparse, and indeed even difficult to define for complex data structures (e.g., for graphs featuring multi-arm trials or uncommon baseline treatments).

The previous example examined a situation where a meta-analysis was planned retrospectively to examine a key safety issue. In the development of medical devices, further examples exist where Bayesian meta-analytic techniques have been used prospectively as way when planning a new study [74]. In the planning stage of one trial, which focused on an adjunctive to percutaneous coronary intervention and stenting, the prior information from previous study was synthesized and then incorporated in the pre-specified analysis plan for the new study (see [75] for further background). More details on this important topic can be found in the companion Drug Information Association Bayesian working group paper that focuses on the design and analysis of safety trials [76].

One final topic for discussion and future research involves extensions that borrow strength across multiple outcomes [29, 77], subgroups [37], or some combination of both [30]. The use of appropriate Bayesian exchangeability or partial exchangeability assumptions seem naturally suited to such problems. However, because of the flexibility of MCMC–Bayesian methods, the number of possible models and structures is vast. Clear guidance on how such approaches should be used in practice thus remains a current research gap. Future work in this area might also link well with the recent FDA guidance on evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes [78], where sponsors are encouraged to perform a meta-analysis of the important cardiovascular events across phase II and III controlled clinical trials and explore similarities and differences across subgroups.

REFERENCES

[1] Huque M. Experiences with meta-analysis in NDA submissions. *Proceedings of the Biopharmaceutical Section American Statistical Association* 1988; **2**:28–33.

[2] Jüni P, Nartey L, Reichenbach S, Sterchi R, Dieppe P, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *The Lancet* 2004; **364**(9450):2021–2029. DOI: 10.1016/S0140-6736(04)17514-4.

[3] Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine* 2007; **356**(24):2457–2471. DOI: 10.1056/NEJMoa072761.

[4] Trelle S, Reichenbach S, Wandel S, Hildebrand P, Tschannen B, Villiger P, Egger M, Jüni P. Cardiovascular safety of non-steroidal anti-inflammatory drugs: network meta-analysis. *BMJ (Clinical research edition)* 2011; **342**:c7086.

[5] Lumley T. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* 2002; **21**(16):2313–2324.

[6] Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* 2004; **23**(20):3105–3124.

[7] Dias S, Welton N, Sutton A, Ades A. NICE DSU Technical Support Document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials 2011. Available at: http://www.nicedsu.org.uk(accessed09.27.2012).

[8] Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, Boersma C, Thompson D, Larholt KM, Diaz M, Barrett A. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR task force on indirect treatment comparisons good research practices: part 2. *Value in Health* 2011; **14**(4):429–437.

[9] Carlin B, Louis T. *Bayesian methods for data analysis*, 3rd ed. Chapman and Hall / CRC Press: Boca Raton, FL, 2009.

[10] Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Wiley: Chichester, 2004.

[11] Senn S. The many modes of meta. *Drug Information Journal* 2000; **34**(2):535–549.

[12] Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B* 2002; **64**:1–34.

[13] Gelfand A, Ghosh S. Model choice: a minimum posterior predictive loss approach. *Biometrika* 1998; **85**:1–11.

[14] Raftery A, Madigan D, Hoeting J. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 1997; **92**:179–191.

[15] Carlin J. Meta-analysis for 2× 2 tables: a Bayesian approach. *Statistics in Medicine* 1992; **11**(2):141–158. DOI: 10.1002/sim.4780110202.

[16] Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**(24):2685–2699. DOI: 10.1002/sim.4780142408.

[17] Higgins J, Thompson S, Spiegelhalter D. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; **172**(1):137–159. DOI: 10.1111/j.1467-985X.2008.00552.x.

[18] Sutton A, Abrams K. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 2001; **10**(4):277–303. DOI: 10.1177/096228020101000404.

[19] Warn D, Thompson S, Spiegelhalter D. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in Medicine* 2002; **21**(11):1601–1623. DOI: 10.1002/sim.1189.

[20] Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* 2005; **24**(15):2401–2428. DOI: 10.1002/sim.2112.

[21] Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *International journal of epidemiology* 2012; **41**(3):818–827.

[22] Sutton A, Kendrick D, Coupland C. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine* 2008; **27**(5):651–669. DOI: 10.1002/sim.2916.

[23] Turner R, Spiegelhalter D, Smith G, Thompson S. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A(Statistics in Society)* 2009; **172**:21–47.

[24] Lee K, Thompson S. Flexible parametric models for random-effects distributions. *Statistics in Medicine* 2007; **27**(3):418–434.

[25] Muthukumarana S, Tiwari RC. Meta-analysis using Dirichlet process. *Statistical Methods in Medical Research* 2012. DOI: 10.1177/0962280212453891. (in press).

[26] Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 2011; **8**(2):129–143. DOI: 10.1177/1740774510396933.

[27] Browne W, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* 2006; **1**(3):473–514.

[28] Daniels M, Hughes M. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**(17):1965–1982.

[29] Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* 2004; **60**(2):418–426. DOI: 10.1111/j.0006-341X.2004.00186.x. PMID: 15180667.

[30] Dumouchel W. Multivariate Bayesian logistic regression for analysis of clinical study safety issues. *Statistical Science* 2012; **27**(3):319–339. DOI: 10.1214/11-STS381.

[31] Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research* 2012. DOI: 10.1177/0962280211432219.

[32] Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006; **101**(474):447–459.

[33] Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* 2009; **10**(4):792–805.

[34] Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991; **10**(11):1665–1677. DOI: 10.1002/sim.4780101105.

[35] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; **23**(9):1351–1375. DOI: 10.1002/sim.1761.

[36] Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* 2007; **26**(1):53–77. DOI: 10.1002/sim.2528.

[37] Jones B, Roger J, Lane P, Lawton A, Fletcher C, Cappelleri J, Tate H, Moneuse P. Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical Statistics* 2011; **10**(6):523–531.

[38] Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**(24):2273–2284. DOI: 10.1002/sim.4780122405.

[39] Piepho HP, Williams ER, Madden LV. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics* 2012; **68**(4):1269–1277. DOI: 10.1111/j.1541-0420.2012.01786.x. (in press).

[40] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* 2006; **1**(3):515–534. DOI: 10.1214/06-BA117A.

[41] Salanti G, Higgins J, Ades A, Ioannidis J. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* 2008; **17**(3):279–301.

[42] Hong H, Carlin B, Shamliyan T, Wyman J, Ramakrishnan R, Sainfort F, Kane R. Comparing Bayesian and frequentist approaches for multiple outcome mixed treatment comparisons. *Medical Decision Making* 2013; **33**(5):702–714.

[43] Dias S, Welton N, Sutton A, Ades A. NICE DSU technical support document 5: evidence synthesis in the baseline natural history model, 2011. Available at: http://www.nicedsu.org.uk(accessed09.27.2012).

[44] Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clinical Trials* 2010; **7**(1):5–18.

[45] O' Hagan A, Buck CE, Daneshkhah A, Eiser JE, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T. *Uncertain judgements: eliciting expert probabilities*. Wiley: Chichester, 2006.

[46] Oakley JE, OHagan A. SHELF: the Sheffield elicitation framework (version 2.0), 2010. Available at: http://www.jeremy-oakley.staff.shef.ac.uk/(accessed04.17.2012).

[47] Kinnersley N, Day S. Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study. *Pharmaceutical statistics* 2013; **12**(2):104–113. DOI: 10.1002/pst.1552.

[48] Little R, Rubin D. *Statistical analysis with missing data*, 2nd. Wiley: New York, 2002.

[49] Hong H, Chu H, Zhang J, Carlin B. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Technical Report 2012-018*, Division of Biostatistics, University of Minnesota, 2012.

[50] Moher D, Liberati A, Tetzlaff J, Altman DG, for the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; **339**(jul21 1):b2535–b2535. DOI: 10.1136/bmj.b2535.

[51] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology* 2009; **62**(10):1006–1012. DOI: 10.1016/j.jclinepi.2009.06.005.

[52] Higgins JPT, Green S (eds). *Cochrane handbook for systematic reviews of interventions version 5.1.0*. The Cochrane Collaboration, 2011. Available at: http://www.cochrane-handbook.org/(accessed 09.27.2012).

[53] Ades A, Caldwell D, Reken S, Welton N, Sutton A, Dias S. NICE DSU technical support document 7: evidence synthesis of treatment efficacy in decision making: a reviewers checklist, 2012. Available at: http://www.nicedsu.org.uk(accessed09.27.2012).

[54] Lang T, Secic M. Considering prior probabilities: reporting Bayesian statistical analyses. *How to report statistics in medicine*. American College of Physicians: Philadelphia, USA, 1997; 231–235.

[55] Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *Journal of Clinical Epidemiology* 2005; **58**(3):261–268. DOI: 10.1016/j.jclinepi.2004.08.010.

[56] Moher D, Schulz K, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology* 2001; **1**(1):1–7. DOI: 10.1186/1471-2288-1-2.

[57] McGettigan P, Henry D. Cardiovascular risk and inhibition of cyclooxygenase: a systematic review of the observational studies of selective and nonselective inhibitors of cyclooxygenase 2. *Journal of the American Medical Association* 2006; **296**(13):1633–1644. DOI: 10.1001/jama.296.13.jrv60011. PMID: 16968831.

[58] Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, Lines C, Riddell R, Morton D, Lanas A, Konstam MA, Baron JA. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *New England Journal of Medicine* 2005; **352**(11):1092–1102. DOI: 10.1056/NEJMoa050493.

[59] Solomon DH, Glynn RJ, Levin R, Avorn J. Nonsteroidal anti-inflammatory drug use and acute myocardial infarction. *Archives of Internal Medicine* 2002; **162**(10):1099–1104.

[60] Whelton A, White WB, Bello AE, Puma JA, Fort JG. Effects of celecoxib and rofecoxib on blood pressure and edema in patients > or = 65 years of age with systemic hypertension and osteoarthritis. *The American Journal of Cardiology* 2002; **90**(9):959–963.

[61] Ray WA, Stein CM, Daugherty JR, Hall K, Arbogast PG, Griffin MR. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet* 2002; **360**(9339):1071–1073. DOI: 10.1016/S0140-6736(02)11131-7. PMID: 12383990.

[62] Mamdani M, Rochon P, Juurlink DN, Anderson GM, Kopp A, Naglie G, Austin PC, Laupacis A. Effect of selective cyclooxygenase 2 inhibitors and naproxen on short-term risk of acute myocardial infarction in the elderly. *Archives of Internal Medicine* 2003; **163**(4):481–486. DOI: 10.1001/archinte.163.4.481.

[63] Mamdani M, Juurlink DN, Lee DS, Rochon PA, Kopp A, Naglie G, Austin PC, Laupacis A, Stukel TA. Cyclo-oxygenase-2 inhibitors versus non-selective non-steroidal anti-inflammatory drugs and congestive heart failure outcomes in elderly patients: a population-based cohort study. *Lancet* 2004; **363**(9423):1751–1756. DOI: 10.1016/S0140-6736(04)16299-5. PMID: 15172772.

[64] Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, Shoor S, Ray WA. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *The Lancet* 2005; **365**(9458):475–481. DOI: 10.1016/S0140-6736(05)17864-7.

[65] McGettigan P, Henry D. Cardiovascular risk with non-steroidal anti-inflammatory drugs: systematic review of population-based controlled observational studies. *PLoS Medicine* 2011; **8**(9):1–18. DOI: 10.1371/journal.pmed.1001098.

[66] Varas-Lorenzo C, Riera-Guardia N, Calingaert B, Castellsague J, Pariente A, Scotti L, Sturkenboom M, Perez-Gutthann S. Stroke risk and NSAIDs: a systematic review of observational studies. *Pharmacoepidemiology and Drug Safety* 2011; **20**(12):1225–1236. DOI: 10.1002/pds.2227.

[67] Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**(4):325–337.

[68] Fu H, Price K, Nilsson M, Ruberg S. Adverse events dose-response relationships via Bayesian indirect and mixed treatment comparison. *Journal of Biopharmaceutical Statistics* 2013; **23**(1):26–42.

[69] Lyman G, Kuderer N. The strengths and limitations of meta-analyses based on aggregate data. *BMC Medical Research Methodology* 2005; **5**(1):1–7.

[70] Berlin JA, Crowe BJ, Whalen E, Xia HA, Koro CE, Kuebler J. Meta-analysis of clinical trial safety data in a drug development program: answers to frequently asked questions. *Clinical Trials* 2013; **10**(1):20–31. DOI: 10.1177/1740774512465495.

[71] Askling J, Fahrbach K, Nordstrom B, Ross S, Schmid C, Symmons D. Cancer risk with tumor necrosis factor alpha (TNF) inhibitors: meta-analysis of randomized controlled trials of adalimumab, etanercept, and infliximab using patient level data. *Pharmacoepidemiology and Drug Safety* 2011; **20**(2):119–130. DOI: 10.1002/pds.2046.

[72] Dias S, Welton N, Caldwell D, Ades A. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* 2010; **29**:932–944. DOI: 10.1002/sim.3767.

[73] Presanis A, Ohlssen D, Spiegelhalter D, DeAngelis D. Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science* 2013. (in press).

[74] Campbell G. Bayesian statistics in medical devices: innovation sparked by the FDA. *Journal of Biopharmaceutical statistics* 2011; **21**(5):871–887. DOI: 10.1080/10543406.2011.589638.

[75] Bonangelino P, Irony T, Liang S, Li X, Mukhi V, Ruan S, Xu Y, Yang X, Wang C. Bayesian approaches in medical device clinical trials: a discussion with examples in the regulatory setting. *Journal of Biopharmaceutical Statistics* 2011; **21**(5):938–953. DOI: 10.1080/10543406.2011.589650.

[76] Price KL, Xia HA, Lakshminarayanan M, Madigan D, Manner D, Scott J, Stamey J, Thompson L. Bayesian methods for design and analysis of safety trials. *Pharmaceutical Statistics* 2014; **13**(1): 13–24.

[77] Xia A, Ma H, Carlin BP. Bayesian hierarchical modeling for detecting safety signals in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **21**(5):1006–1029. DOI: 10.1080/10543406.2010.520181.

[78] U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). Guidance for industry diabetes mellitus - evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes, 2008.